W728–W733 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki415

The multilocus sequence typing network: mlst.net

David M. Aanensen* and Brian G. Spratt

Department of Infectious Disease Epidemiology, Imperial College London, St Mary's Hospital, London W2 1PG, UK

Received February 14, 2005; Revised and Accepted March 21, 2005

ABSTRACT

The unambiguous characterization of strains of a pathogen is crucial for addressing questions relating to its epidemiology, population and evolutionary biology. Multilocus sequence typing (MLST), which defines strains from the sequences at seven house-keeping loci, has become the method of choice for molecular typing of many bacterial and fungal pathogens (and non-pathogens), and MLST schemes and strain databases are available for a growing number of prokaryotic and eukaryotic organisms. Sequence data are ideal for strain characterization as they are unambiguous, meaning strains can readily be compared between laboratories via the Internet. Laboratories undertaking MLST can quickly progress from sequencing the seven gene fragments to characterizing their strains and relating them to those submitted by others and to the population as a whole. We provide the gateway to a number of MLST schemes, each of which contain a set of tools for the initial characterization of strains, and methods for relating query strains to other strains of the species, including clustering based on differences in allelic profiles, phylogenetic trees based on concatenated sequences, and a recently developed method (eBURST) for identifying clonal complexes within a species and displaying the overall structure of the population. This network of MLST websites is available at http://www.mlst.net

INTRODUCTION

Multilocus sequence typing (MLST) is a nucleotide sequence-based approach to the unambiguous characterization of strains of bacterial species, or other microbial species, via the Internet (1,2). MLST involves obtaining the sequences of internal fragments of seven house-keeping genes for each strain of a particular species. The sequences of each fragment are compared with all the previously identified sequences (alleles) at that locus and, thereby, are assigned allele numbers at each of the seven loci. The combination of the seven allele numbers defines the allelic profile of the strain and each different allelic profile is assigned as a sequence type (ST), which is used to describe the strain.

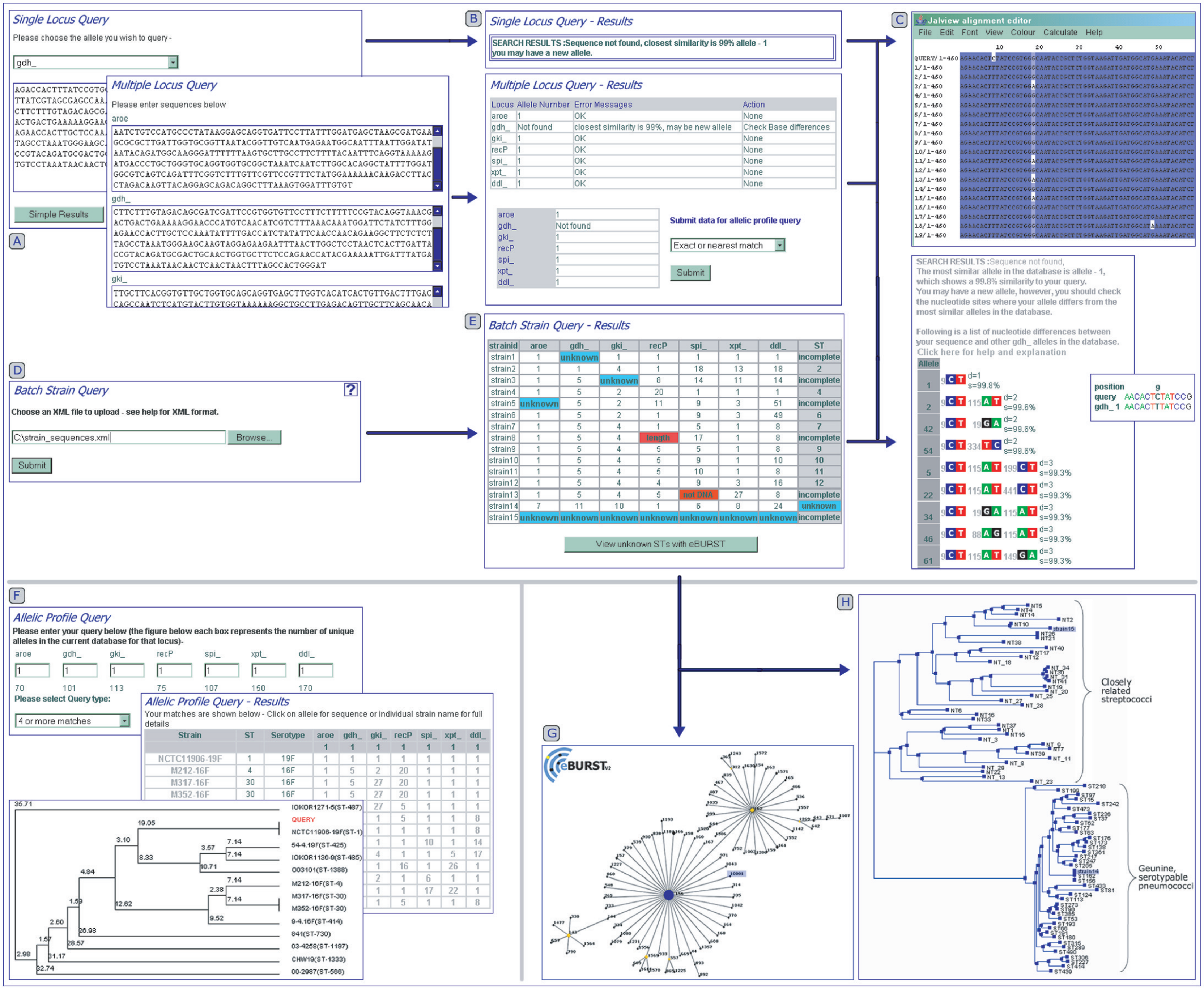
Nucleotide sequencing is relatively cheap, and easy to perform. The data produced by MLST are ideal for the characterization of strains of bacterial or fungal species via a web server. MLST is now widely used for molecular epidemiology as it allows strains studied by different groups to be compared and MLST schemes have been developed for ∼20 bacteria (mostly pathogens) (3), and three fungi (4,5) and databases that can be queried have been available for several years (6). The MLST databases are currently hosted on two main web servers located at Imperial College London (http://www.mlst.net) and Oxford University [http://pubmlst.org; (7)]. The former web server acts as a gateway to a number of species-specific websites each of which contains tools for the analysis of allele sequences and STs, and a web interface for obtaining epidemiological information held on the increasing numbers of strains that are submitted by the user community.

Along with centrally available tools for those interested in starting their own MLST schemes, such as for defining alleles using non-redundant databases (NRDB), measuring linkage disequilibrium and an interface to Splits Tree (8), http://www.mlst.net provides a number of options to display the relatedness of query strains to those in the strain database.

MLST WORKFLOW

Laboratories undertaking MLST can access species-specific information on each of the individual mlst.net species websites, including sequencing protocols and primer sequences, allowing a laboratory to begin producing data rapidly. Characterization of a strain requires the generation of the sequences of the seven gene fragments and, once these are available, they are used to query the appropriate mlst.net website, to assign the alleles at each locus and thereby to obtain the allelic profile of the query strain. Each MLST website holds the sequences of all known alleles at each of the seven loci, and all known allelic profiles, and through the curator assigns new allele numbers and STs. Every different sequence at each locus is assigned as a distinct allele and new alleles are assigned allele numbers by the curator and are entered in the allele database.

*To whom correspondence should be addressed. Tel: +44 0 20 7594 3825; Fax: +44 0 20 7594 3693; Email: d.aanensen@imperial.ac.uk



Single Locus Query

Please choose the allele you wish to query -

gdh_

```
AGACCACTTTATCCGTG
TTATCGTAGCGAGCCAA
CTTCTTTGTAGACAGCG
ACTGACTGAAAAAGGAA
AGAACCACTTGCTCCAA
TAGCCTAAATGGGAAGC
CCGTACAGATGCGACTG
TGTCCTAAATAACAACT
```

Multiple Locus Query

Please enter sequences below

aroe

```
AATCTGTCCATGCCCTATAAGGAGCAGGTGATTCCTTATTTGGATGAGCTAAGCGATGAA
GCGCGCTTGATTGGTGCGGTTAATACGGTTGTCAATGAGAATGGCAATTTAATTGGATAT
AATACAGATGGCAAGGGATTTTTTAAGTGCTTGCCTTCTTTTACAATTTCAGGTAAAAAG
ATGACCCTGCTGGGTGCAGGTGGTGCGGCTAAATCAATCTTGGCACAGGCTATTTTGGAT
GGCGTCAGTCAGATTTCGGTCTTTGTTCGTTCCGTTTCTATGGAAAAAACAAGACCTTAC
CTAGACAAGTTACAGGAGCAGACAGGCTTTAAAGTGGATTTGTGT
```

gdh_

```
CTTCTTTGTAGACAGCGATCGATTCCGTGGTGTTCCTTTCTTTTTCCGTACAGGTAAACG
ACTGACTGAAAAAGGAACCCATGTCAACATCGTCTTTAAACAAATGGATTCTATCTTTGG
AGAACCACTTGCTCCAAATATTTTGACCATCTATATTCAACCAACAGAAGGCTTCTCTCT
TAGCCTAAATGGGAAGCAAGTAGGAGAAGAATTTAACTTGGCTCCTAACTCACTTGATTA
CCGTACAGATGCGACTGCAACTGGTGCTTCTCCAGAACCATACGAAAAAATTGATTTATGA
TGTCCTAAATAACAACTCAACTAACTTTAGCCACTGGGAT
```

gki_

```
TTGCTTCACGGTGTTGCTGGTGCAGCAGGTGAGCTTGGTCACATCACTGTTGACTTTGAC
CAGCCAATCTCATGTACTTGTGGTAAAAAAGGCTGCCTTGAGACAGTTGCTTCAGCAACA
```

Single Locus Query - Results

SEARCH RESULTS :Sequence not found, closest similarity is 99% allele - 1
you may have a new allele.

Multiple Locus Query - Results

Locus	Allele Number	Error Messages	Action
aroe	1	OK	None
gdh_	Not found	closest similarity is 99%, may be new allele	Check Base differences
gki_	1	OK	None
recP	1	OK	None
spi_	1	OK	None
xpt_	1	OK	None
ddl_	1	OK	None

aroe	1
gdh_	Not found
gki_	1
recP	1
spi_	1
xpt_	1
ddl_	1

Submit data for allelic profile query

Exact or nearest match

```
             10        20        30        40        50
QUERY/1-460 AGAACACTCTATCCGTGGGCAATACCGCTCTGGTAAGATTGATGGCATGAAATACATCT
1/1-460     AGAACACTTTATCCGTGGGCAATACCGCTCTGGTAAGATTGATGGCATGAAATACATCT
2/1-460     AGAACACTTTATCCGTGGGCAATACCGCTCTGGTAAGATTGATGGCATGAAATACATCT
3/1-460     AGAACACTTTATCCGTGGACAATACCGCTCTGGTAAGATTGATGGCATGAAATACATCT
4/1-460     AGAACACTTTATCCGTGGGCAATACCGCTCTGGTAAGATTGATGGCATGAAATACATCT
5/1-460     AGAACACTTTATCCGTGGGCAATACCGCTCTGGTAAGATTGATGGCATGAAATACATCT
6/1-460     AGAACACTTTATCCGTGGGCAATACCGCTCTGGTAAGATTGATGGCATGAAATACATCT
7/1-460     AGAACACTTTATCCGTGGGCAATACCGCTCTGGTAAGATTGATGGCATGAAATACATCT
8/1-460     AGAACACTTTATCCGTGGGCAATACCGCTCTGGTAAGATTGATGGCATGAAATACATCT
9/1-460     AGAACACTTTATCCGTGGGCAATACCGCTCTGGTAAGATTGATGGCATGAAATACATCT
10/1-460    AGAACACTTTATCCGTGGGCAATACCGCTCTGGTAAGATTGATGGCATGAAATACATCT
11/1-460    AGAACACTTTATCCGTGGACAATACCGCTCTGGTAAGATTGATGGCATGAAATACATCT
12/1-460    AGAACACTTTATCCGTGGGCAATACCGCTCTGGTAAGATTGATGGCATGAAATACATCT
13/1-460    AGAACACTTTATCCGTGGACAATACCGCTCTGGTAAGATTGATGGCATGAAATACATCT
14/1-460    AGAACACTTTATCCGTGGGCAATACCGCTCTGGTAAGATTGATGGCATGAAATACATCT
15/1-460    AGAACACTTTATCCGTGGACAATACCGCTCTGGTAAGATTGATGGCATGAAATACATCT
16/1-460    AGAACACTTTATCCGTGGGCAATACCGCTCTGGTAAGATTGATGGCATGAAATACATCT
17/1-460    AGAACACTTTATCCGTGGGCAATACCGCTCTGGTAAGATTGATGGCATGAAATACATCT
18/1-460    AGAACACTTTATCCGTGGGCAATACCGCTCTGGTAAGATTGATGGCATAAAATACATCT
19/1-460    AGAACACTTTATCCGTGGGCAATACCGCTCTGGTAAGATTGATGGCATGAAATACATCT
```

SEARCH RESULTS :Sequence not found,
The most similar allele in the database is allele - 1, which shows a 99.8% similarity to your query.
You may have a new allele, however, you should check the nucleotide sites where your allele differs from the most similar alleles in the database.

Following is a list of nucleotide differences between your sequence and other gdh_ alleles in the database.
Click here for help and explanation

Allele				

1	9 C T			d=1
				s=99.8%

2	9 C T	115 A T		d=2
				s=99.6%

42	9 C T	19 G A		d=2
				s=99.6%

54	9 C T	334 T C		d=2
				s=99.6%

5	9 C T	115 A T	199 C T	d=3
				s=99.3%

22	9 C T	115 A T	441 C T	d=3
				s=99.3%

34	9 C T	19 G A	115 A T	d=3
				s=99.3%

46	9 C T	88 A G	115 A T	d=3
				s=99.3%

61	9 C T	115 A T	149 G A	d=3
				s=99.3%

```
position       9
query   AACACTCTATCCG
gdh_1   AACACTTTATCCG
```

Batch Strain Query

Choose an XML file to upload - see help for XML format.

C:\strain_sequences.xml

Batch Strain Query - Results

strainid	aroe	gdh_	gki_	recP	spi_	xpt_	ddl_	ST
strain1	1	unknown	1	1	1	1	1	incomplete
strain2	1	1	4	1	18	13	18	2
strain3	1	5	unknown	8	14	11	14	incomplete
strain4	1	5	2	20	1	1	1	4
strain5	unknown	5	2	11	9	3	51	incomplete
strain6	1	5	2	1	9	3	49	6
strain7	1	5	4	1	5	1	8	7
strain8	1	5	4	length	17	1	8	incomplete
strain9	1	5	4	5	5	1	8	9
strain10	1	5	4	5	9	1	10	10
strain11	1	5	4	5	10	1	8	11
strain12	1	5	4	4	9	3	16	12
strain13	1	5	4	5	not DNA	27	8	incomplete
strain14	7	11	10	1	6	8	24	unknown
strain15	unknown	unknown	unknown	unknown	unknown	unknown	unknown	incomplete

Allelic Profile Query

Please enter your query below (the figure below each box represents the number of unique alleles in the current database for that locus)-

aroe	gdh_	gki_	recP	spi_	xpt_	ddl_
1	1	1	1	1	1	1
70	101	113	75	107	150	170

Please select Query type:

4 or more matches

Allelic Profile Query - Results

Your matches are shown below - Click on allele for sequence or individual strain name for full details

Strain	ST	Serotype	aroe	gdh_	gki_	recP	spi_	xpt_	ddl_
			1	1	1	1	1	1	1
NCTC11906-19F	1	19F	1	1	1	1	1	1	1
M212-16F	4	16F	1	5	2	20	1	1	1
M317-16F	30	16F	1	5	27	20	1	1	1
M352-16F	30	16F	1	5	27	20	1	1	1
					27	5	1	1	1
					1	5	1	1	8
					1	1	1	1	8
					1	1	10	1	14
					4	1	1	5	17
					1	16	1	26	1
					2	1	6	1	1
					1	1	17	22	1
					1	5	1	1	8

Each MLST species website offers a number of analysis steps for a user. First, alleles have to be assigned from the sequence data by one of three options (Figure 1A and D):

Single/batch locus query: allowing a single sequence or a batch of sequences for a single locus to be compared with all known alleles.

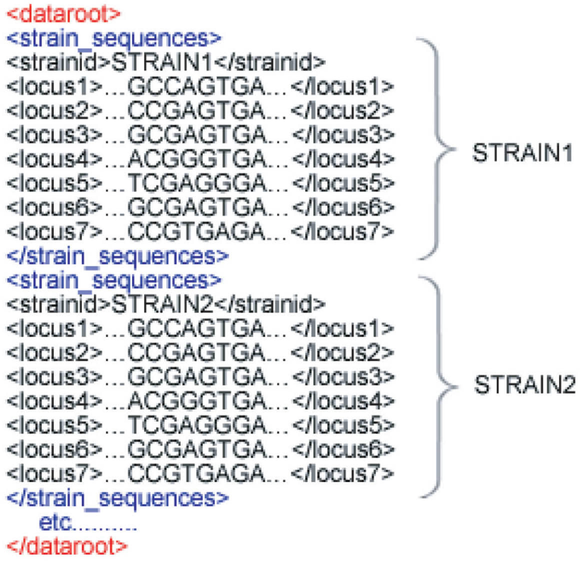
Multiple locus query: allowing the input of the sequences of all seven loci for a single strain.

Batch strain query: allowing input of the sequences of all seven loci for a batch of strains.

In all cases, the user's sequence is checked for correct length for that locus, and for the absence of unexpected characters, and is then queried against all other sequences in the species database. For *Candida albicans*, a diploid organism, the standard ambiguity codes are allowed and are used to assign heterozygous nucleotide sites (4).

If the user's sequence is found, the allele number is returned, whereas if the user has a novel sequence, the percentage identity to the closest allele in the database is returned and the user is advised to check carefully those nucleotide sites that differ from the most similar allele or alleles in the database (Figure 1B). This can be carried out using the Jalview alignment editor (9), or the nucleotide differences can be displayed between the query sequence and the most similar alleles, as in Figure 1C. The latter method allows the user to view the flanking sequence around each nucleotide difference between the query and the most similar alleles, allowing the trace files of their proposed new allele to be searched easily for any potential ambiguities or sequencing errors. If a user is confident that they have a new allele, the forward and reverse trace files are submitted to the MLST scheme curator, as a quality control check, before a new allele number is assigned by the curator for the novel sequence. The sequence of the new allele is then entered into the database.

Repeating this process for each locus provides the seven-digit allelic profile for the query strain. The seven-digit allelic profile can then be entered into the allelic profile query to discover if the strain is identical, or similar, in allelic profile to any of the strains already in the database (Figure 1F). The multiple locus query represents a batch processing method for a single strain, allowing all seven sequences for the query strain to be entered at once and for the allele numbers at each locus to be returned.



```
<dataroot>
<strain_sequences>
<strainid>STRAIN1</strainid>
<locus1>...GCCAGTGA...</locus1>
<locus2>...CCGAGTGA...</locus2>
<locus3>...GCGAGTGA...</locus3>
<locus4>...ACGGGTGA...</locus4>
<locus5>...TCGAGGGA...</locus5>
<locus6>...GCGAGTGA...</locus6>
<locus7>...CCGTGAGA...</locus7>
</strain_sequences>
<strain_sequences>
<strainid>STRAIN2</strainid>
<locus1>...GCCAGTGA...</locus1>
<locus2>...CCGAGTGA...</locus2>
<locus3>...GCGAGTGA...</locus3>
<locus4>...ACGGGTGA...</locus4>
<locus5>...TCGAGGGA...</locus5>
<locus6>...GCGAGTGA...</locus6>
<locus7>...CCGTGAGA...</locus7>
</strain_sequences>
   etc..........
</dataroot>
```

STRAIN1
STRAIN2

Figure 2. The XML format for batch querying multiple strains using mlst.net.

BATCH STRAIN QUERY

The repeated querying of single strains becomes very time consuming for laboratories undertaking MLST on many hundred strains of a particular species. The need to analyze sequences from multiple genes from a large number of strains at the same time precludes the use of standard sequence formats such as FASTA or MEGA. Therefore, we use a simple XML format that allows the batch processing of hundreds of strains at one time (Figure 2).

Formatting the input data with a basic XML wrapper around a set of seven sequences for each strain allows a user to produce a file, for an unlimited number of strains, that can be used for batch processing. To aid production of such a format, each of the MLST species subsites at http://www.mlst.net provides a modified Access database that allows users to store their sequence and strain information in one place, and allows the data to be exported in bulk in the correct XML format, without the need for a user to manually produce the document. Furthermore, for sequencing laboratories using the STARS (http://www.molbiol.ox.ac.uk/~paediat/stars/) platform for MLST, we provide a facility to convert the FASTA files generated by STARS into the XML format via a web form.

When a user uploads the generated XML file (Figure 1D), the sequences for each of the seven loci in all of the strains are checked for invalid characters and correct length. Each sequence is then queried against the appropriate allele database (Figure 1E). If found, the allele number is returned and, if unknown, the user can look further into the sequence differences between the query allele and the most similar alleles in the database (Figure 1C). If all the seven loci are found, the allelic profile of the strain is queried against a look-up table of STs within the database and, if a match is found, the ST number is returned. If the allelic profile is previously unknown this information is returned. The batch procedure, therefore, automatically returns a table with the alleles, allelic profiles

Figure 1. Schematic representation of a typical MLST workflow. (**A**) Sequences can be entered locus by locus (single locus query), or all seven loci from a single strain (multiple locus query), or (D) by uploading a XML file with a set of strains and their sequences (batch strain query). (**B**) For the single and multiple locus queries, sequences that are not in the database are identified, and can be compared with the sequences of all known alleles using Jalview (7), or (**C**) the nucleotide differences compared with the most similar alleles can be displayed. The batch strain query (**D**) returns a strain table (**E**), which shows the allele number for each locus if known and the allelic profile if all the seven alleles and the ST are known. Strains that have the most similar allelic profiles to query strains are displayed as a table or by cluster analysis (**F**), and further information about them can be obtained. For the pneumococcal example used here, the query strains can be compared with the reference set of pneumococcal strains and closely related streptococcal strains, to establish whether or not they are pneumococci, using the concatenated sequences to construct a neighbor-joining tree (**H**). The relationship of unknown strains to the whole population can also be investigated using eBURST (**G**).

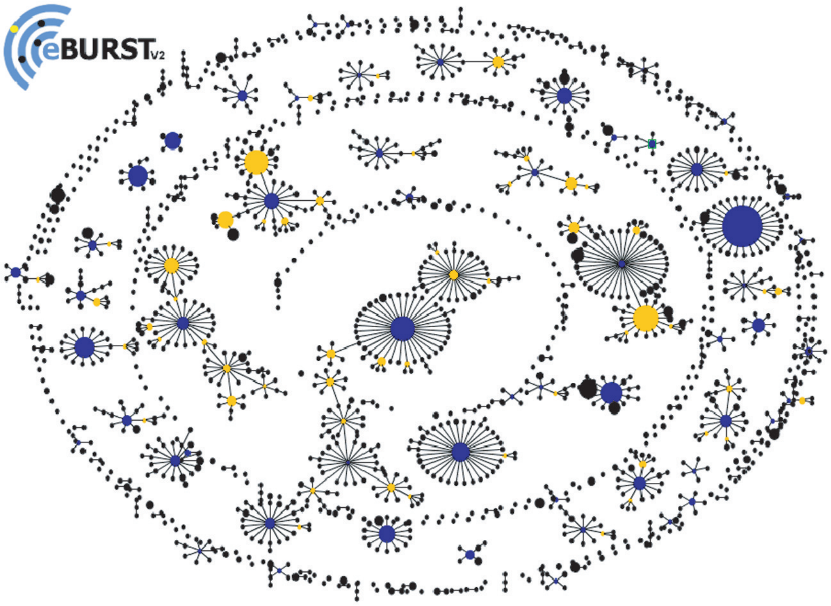


Figure 3. A population snapshot of the entire *S.pneumoniae* MLST database showing all major and minor clonal complexes viewed using eBURST.

and STs of all the input strains, flagging up those alleles and STs that are previously unknown (Figure 1E).

Comparing query strains to the database: clustering using allelic profiles

The simplest approach is to identify those strains in the database that have some minimum level of similarity in their allelic profile to each query strain (e.g. sharing alleles at ⩾4 of the seven loci), and to show the relationship of the query strain to those returned from the database query using a dendrogram, based on the matrix of pairwise differences between the allelic profiles of the strains (Figure 1F).

Comparing query strains to the database: using eBURST

Traditionally, dendrograms have been the method of choice for displaying the implied relationships between strains of a bacterial population or species. However, although dendrograms are good at visualizing the clusters of identical or very similar strains, the bifurcating process of lineage splitting implied by a dendrogram is a very poor representation of the way in which bacterial lineages emerge and diversify. A new algorithm, BURST (10), was recently introduced that does not impose a tree-like pattern of descent, but rather uses an appropriate model of recent bacterial evolution. In addition, it is very difficult to display the relatedness of all strains in a large MLST database, including thousands of STs, on a dendrogram, and better ways of displaying the relationships among all strains in large MLST databases are required.

Briefly, the model incorporated into BURST assumes that, due to selection or genetic drift, some genotypes will occasionally increase in frequency in the population and will then gradually diversify by the accumulation of mutation(s) and/or recombinational replacements, resulting in slight variants of the founding genotype. Initially, members of this emerging clone will be indistinguishable in allelic profile by MLST, however with time, the clone will diversify to produce a number of variants in which one of the seven MLST loci has been altered—single locus variants (SLVs). Further diversification will produce variants of the founder ST that differ at two out of the seven loci—double locus variants (DLVs). In this simple model, bacterial populations will consist of a series of clonal complexes (sets of variants of a founding genotype) that can be recognized from the allelic profiles of the strains within a MLST database (10).

An interactive implementation of the BURST algorithm, eBURSTv2 (10), is integrated within the MLST websites at http://www.mlst.net as a JAVA^{TM} applet and can be used to explore the relationships among strains within the database and to explore the relationships of newly characterized strains to those in the database (Figure 1G). eBURST uses the STs and their associated allelic profiles as input and, using the default setting, divides the strains into groups in which all STs in the same group share ⩾6 out of 7 loci with at least one other member of that group, resulting in non-overlapping groups or clonal complexes. Of particular value is the ability to link back to the MLST database from the eBURST diagram of a clonal complex, and the ability to display all the STs in a large MLST database in a single diagram [(10); a population snapshot; and Figure 3], showing all the major and minor clonal complexes, and individual STs that are relatively distantly related to all other STs.

Comparing query strains to the database: using the concatenated sequences

The ability to concatenate the sequences at the seven loci, maintaining the correct reading frame, and to construct a neighbor-joining tree based on these sequences is provided, but needs to be used with considerable caution. A module from MEGA (11) provides the tree topology in Newick format which is then displayed using the ATV applet (12). Allelic changes at the MLST loci will occur (to a varying degree depending on the species) by recombination, and in many cases the relative contribution of recombination and point mutation to the diversification of strains will be unknown (13). A long history of recombination will preclude the recovery of the true phylogenetic relationships between distantly related bacterial strains and even the relatedness between similar strains may be better represented on a tree based on differences in allelic profiles than one based on differences in the concatenated sequences. However, there are specific issues that can be usefully addressed by using the concatenated sequences. For example, the *Burkholderia pseudomallei* database includes strains of closely related species and the *B.pseudomallei* MLST website provides a facility to examine the position of a query strain on the tree constructed using concatenated sequences, which can establish whether the query strain is *B.pseudomallei* or something similar to, but distinct from, *B.pseudomallei* (14). Similarly, there is considerable confusion about whether strains that appear to be *Streptococcus pneumoniae*, but which cannot be assigned to a pneumococcal capsular serotype, are authentic pneumococci that do not produce a capsule or are members of a similar but distinct streptococcal population. The pneumococcal MLST website has a facility to examine whether a query strain clusters within a reference set of *S.pneumoniae* strains, or with the related population, using a tree based on concatenated sequences, which can resolve this issue in most cases (see the following section; Figure 1H). Trees based on concatenated sequences may also be useful for assigning *Haemophilus influenzae* strains to major lineages (15) or for *Staphylococcus aureus* where recombination appears to be rare (16).

Typical workflow for data entry using the batch strain query

Here, we consider the workflow of a user analyzing a number of recently sequenced strains using batch entry. As an illustrative example we focus on a single representative mlst.net species website, http://spneumoniae.mlst.net, the site for characterizing strains of *S.pneumoniae* (17).

The uploaded XML file of a batch of *S.pneumoniae* strains and their associated sequences results in a table of results (Figure 1E). Error messages (red) alert the user to the fact that some sequences are of the wrong length for that locus (strain 8) or contain unexpected characters (strain 13). In some strains, all the alleles are previously known and the allele numbers are returned in the results table. For some of these strains, the combinations of alleles at the seven loci (allelic profiles) are also known and the ST number is shown in the table (e.g. strain 4). In one case (strain 14) the alleles are all known but the combination of alleles is previously unknown. In other strains, one or more alleles are unknown and the ST must also be unknown (e.g. strain 3), and the ST is flagged as incomplete, as the new alleles have to be checked and assigned new allele numbers by the curator. Clicking on ‘unknown’ allele highlights the nucleotide differences in the new allele compared with the most similar alleles (Figure 1C).

None of the alleles in strain 15 are found in the *S.pneumoniae* database, and there is therefore some uncertainty whether this strain is a pneumococcus. To investigate the status of this strain further, the user can select the option to examine the phylogenetic status of the strain, by using the concatenated sequences to compare its position on a reference tree (Figure 1H), which includes a set of strains covering the known diversity of authentic pneumococci, and a set of closely related strains that are similar to but distinct from the authentic pneumococci (W. P. Hanage and B. G. Spratt, unpublished data). The sequences of the loci of the query strain are concatenated, and the sequence is added to a stored file containing the concatenated sequences of the reference strains, and a neighbor-joining tree is constructed (Figure 1H). Using this approach, strain 15, which has an unknown allelic profile but known alleles at all loci, clusters within the authentic pneumococci, but strain 14 with new alleles at all loci is clearly not a pneumococcus, as it clusters away from the pneumococci and within the more diverse set of related streptococcal strains.

From the results of the batch strain query, the user can also relate their unknown STs to all other strains in the MLST database using eBURST (Figure 1G). The unknown STs are assigned unique temporary ST numbers, to distinguish them from the STs in the database. In Figure 1G, strain 14 has been assigned the temporary ST10001 and by eBURST it can be seen to be a SLV of ST156 within one of the major pneumococcal clonal complexes. Any strain in the batch strain query (excepting those with alleles of incorrect length or with unexpected characters) can be compared with the MLST database as, using the eBURST option, new alleles, as well as new STs, are given temporary numbers allowing them to be analyzed by the program.

CONCLUSIONS

Websites for evaluating the taxonomic status of strains using 16S rRNA sequences are well established and, in recent years, several websites have been developed for molecular epidemiology and population genetics, to assign isolates of bacterial species to strains, lineages and clonal complexes, using data generated by MLST. We describe the set of MLST species websites within http://www.mlst.net, and the tools that allow users to identify query strains, and to explore their relationship with other strains in the database. MLST is being widely used and there is a need for new ways to input and query large sets of strains and to display the relatedness of the many thousands of strains within the larger MLST websites. Some progress has been made to achieve these aims and in future we envisage a fully automated procedure, with data flowing directly from sequencer to ST assignment. Those developing new MLST schemes for bacterial or fungal species can join http://www.mlst.net to take advantage of the features available at this site, and to have a consistency of format for the MLST websites. A slightly different common format for MLST websites is provided by those species sites (such as that for *Neisseria meningitidis*) hosted at http://pubmlst.org. Hosting of new MLST schemes at http://www.mlst.net allows the databases to be stored and backed up on servers at Imperial College London but with remote strain entry, ownership and curation, by the developer of the MLST scheme, using MLST curation software.

ACKNOWLEDGEMENTS

We acknowledge the contribution of Man-Suen Chan who developed the initial MLST website software, James Abbott for the STARS XML conversion script, and Sudhir Kumar and Joel Dudley for providing the server-side MEGA module for concatenated tree drawing. The development of http://www.mlst.net was supported by the Wellcome Trust. Part of the publication cost for this article was provided by JISC. Funding to pay the Open Access publication charges for this article was provided by the Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Maiden,M.C.J., Bygraves,J.A., Feil,E., Morelli,G., Russell,J.E., Urwin,R., Zhang,Q., Zhou,J., Zurth,K., Caugant,D.A. *et al*. (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl Acad. Sci. USA*, **95**, 3140–3145.
2. Hanage,W.P., Feil,E.J., Brueggemann,A.B. and Spratt,B.G. (2004) Multilocus sequence typing: strain characterization, population biology, and patterns of evolutionary descent. In Persing,D.H., Tenover,F.C., Versalovic,J., Tang,Y., Unger,E.R., Relman,D.A. and White,T.J. (eds), *Molecular Microbiology: Diagnostic Principles and Practice*. American Society Press, Washington DC, pp. 235–243.
3. Urwin,R. and Maiden,M.C. (2003) Multi-locus sequence typing: a tool for global epidemiology. *Trends Microbiol*., **10**, 479–487.
4. Bougnoux,M.E., Aanensen,D.M., Morand,S., Theraud,M., Spratt,B.G. and d'Enfert,C. (2004) Multilocus sequence typing of *Candida albicans*: data exchange and applications. *Infect. Genet. Evol*., **6**, 243–252.
5. Dodgson,A.R., Pujol,C., Denning,D.W., Soll,D.R. and Fox,A.J. (2003) Multilocus sequence typing of *Candida glabrata* reveals geographically enriched clades. *J. Clin. Microbiol*., **12**, 5709–5717.
6. Chan,M.S., Maiden,M.C.J. and Spratt,B.G. (2001) Database-driven multi locus sequence typing (MLST) of bacterial pathogens. *Bioinformatics*, **17**, 1077–1083.
7. Jolley,K.A., Chan,M.S. and Maiden,M.C.J. (2004) mlstdbNet—distributed multi-locus sequence typing (MLST) databases. *BMC Bioinformatics*, **5**, 86.
8. Huson,D.H. (1998) SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*, **14**, 68–73.
9. Clamp,M., Cuff,J., Searle,S.M. and Barton,G.J. (2004) The Jalview Java alignment editor. *Bioinformatics*, **3**, 426–427.
10. Feil,E.J., Li,B.C., Aanensen,D.M., Hanage,W.P. and Spratt,B.G. (2004) eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J. Bacteriol*., **186**, 1518–1530.
11. Kumar,S., Tamura,K., Jakobsen,I.B. and Nei,M. (2001) MEGA2: Molecular Evolutionary Genetics Analysis software. *Bioinformatics*, **17**, 1244–1245.
12. Zmasek,C.M. and Eddy,S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, **17**, 383–384.
13. Feil,E.J., Holmes,E.C., Bessen,D.E., Chan,M.S., Day,N.P., Enright,M.C., Goldstein,R., Hood,D.W., Kalia,A., Moore,C.E. *et al*. (2001) Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proc. Natl Acad. Sci. USA*, **98**, 182–187.
14. Godoy,D., Randle,G., Simpson,A.J., Aanensen,D.M., Pitt,T.L., Kinoshita,R. and Spratt,B.G. (2003) Multilocus sequence typing and evolutionary relationships among the causative agents of melioidosis and glanders, *Burkholderia pseudomallei* and *Burkholderia mallei*. *J. Clin. Microbiol*., **5**, 2068–2079.
15. Meats,E., Feil,E.J., Stringer,S., Cody,A.J., Goldstein,R., Kroll,J.S., Popovic,T. and Spratt,B.G. (2003) Characterization of encapsulated and non-capsulated *Haemophilus influenzae*, and determination of phylogenetic relationships, using multilocus sequence typing. *J. Clin. Microbiol*., **41**, 1623–1636.
16. Feil,E.J., Cooper,J.E., Grundmann,H., Robinson,D.A., Enright,M.C., Berendt,A., Peacock,S., Maynard Smith,J., Murphy,M., Spratt,B.G. *et al*. (2003) How clonal is *Staphylococcus aureus*? *J. Bacteriol*., **185**, 3307–3316.
17. Enright,M.C. and Spratt,B.G. (1998) A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. *Microbiology*, **144**, 349–360.